Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Bachelor Thesis

Bachelor Thesis

Developing a Database Application to Compare the Google Books Ngram Corpus to German News Corpora

In 2010, Google launched the *Google Books Ngram Viewer* (GBNV).[a] It allows users to easily plot relative word frequencies, based on the *Google Books Ngram Corpus* (GBNC). The GBNC is extracted from an unknown subset of all the books scanned for the *Google Books* project – and claims to contain data for 6% of all books ever published [1]. Comparing the usage patterns of words within the GBNC to the ones of other, similar corpora would shed light on the composition of the GBNC. The goal of this thesis is to design and implement a system for such comparisons.

While the GBNC can be a valuable resource for different fields of science, numerous concerns have been raised regarding the conclusions drawn from it. One important problem is the lack of metadata for the corpus. Google has never published its exact composition. This limits the validity of conclusions drawn from the GBNC, for example in sociology. While some studies have addressed this problem [2, 3], a methodology to estimate the impact of lacking metadata is still missing.

A first step of this thesis is the creation of a new corpus based on German news outlets. One can then leverage knowledge about the composition of this corpus to better understand the composition of the GBNC. To do so, a query algebra needs to be developed, together with a system implementing it.

The system *ngramSQL* is being developed at the chair [4]. It currently offers only *intra-corporal* operators, i.e., operators working on single n-gram corpora. *ngramSQL* being based on *Apache Spark SQL* can be extended by adding new operators. In the thesis, *inter-corporal* operators shall be added to *ngramSQL*. An example could be an operator that finds words with vastly different usage patterns in different corpora. Such an operator can be implemented in different ways, for example by measuring correlation or simply absolute distances between frequencies of different words. You will be asked to compare the usefulness of different similarity measures for time series and performance of respective implementations in the special case of n-gram corpora. **The thesis will address the following topics:**

- Literature review on metadata reconstruction and the GBNC,
- Design and implementation of inter-corporal operators as an extension to *ngramSQL*,
- Systematic evaluation of the operators with regard to resource usage and scalability.

Basic programming skills in Python and/or Java are necessary. Knowledge about *Apache Spark SQL* is useful, but not required. During your work, you will have access to the chair's computing infrastructure – as well as support and feedback from your advisor in frequent meetings.

[1] Y. Lin et al. "Syntactic Annotations for the Google Books Ngram Corpus". In: *Proceedings of the ACL 2012 system demonstrations*. 2012, pp. 169–174.

[2] A. Koplenig. "The Impact of Lacking Metadata for the Measurement of Cultural and Linguistic Change Using the Google Ngram Data Sets – Reconstructing the Composition of the German Corpus in Times of WWII". In: *Digital Scholarship in the Humanities* 32.1 (2017), pp. 169–188.

[3] N. Younes and U.-D. Reips. "Guideline for Improving the Reliability of Google Ngram Studies: Evidence from Religious Terms". In: *PloS one* 14.3 (2019), e0213554.

[4] F. Richter and K. Böhm. "Querying the Google Books Ngram Corpus – Language, System and Evaluation". In: *To appear*.

[a] https://books.google.com/ngrams

Ansprechpartner

Fabian Richter	fabian.richter@kit.edu	Raum: 364
Am Fasanengarten 5	76131 Karlsruhe	Gebäude: 50.34